Data Analytics, Autumn 2024
Assignment 1

Recently, former US President and 2024 Republican presidential candidate Donald Trump was shot at an outdoor rally in Pennsylvania on a Saturday evening. In response, the intelligence team conducted a nationwide random sampling of individuals to identify potential suspects. The collected data includes various attributes such as age, gender, and occupation, which are detailed in the provided dataframe (Synthetic Data). The last column of the dataframe indicates the probability of each individual being either a criminal or innocent, with two possible labels: <=0.5 and >0.5. For classification purposes, these labels represent distinct classes: a probability greater than 0.5 suggests a potential criminal, while a probability of 0.5 or less indicates innocence.

Given this dataset, perform the following tasks. You may handle missing values (if any) according to a scheme of your choice.

A. Perform an Exploratory Data Analysis (EDA) on the dataset. EDA may include frequency distribution, univariate and multivariate correlation analysis, as well as basic data visualization (remember the EDA tutorial on 16th August). You are required to prepare a report (Latex Template) in pdf that includes Exploratory Data Analysis (EDA) and insights gained from each implementation.

B. Implement Naive Bayes model on the given dataset without relying on any machine learning libraries (e.g., *sklearn*). Your task is to code the Naive Bayes algorithm from scratch to classify individuals as either criminal or innocent. You may use basic packages such as *numpy*, *pandas*, and *math*. No marks will be given if you use any other custom classifier or ML libraries.

C. Now, implement Naive Bayes, SVM, Decision Tree, and KNN using the *sklearn* module to perform the classification task. Compare the performance of the *sklearn* Naive Bayes implementation with your custom Naive Bayes implementation from part B.

D. Finally, implement an ensemble model by combining multiple classifiers, including your custom Naive Bayes implementation (without *sklearn*), the *sklearn* version of SVM, Decision Tree, and KNN. You must write the ensemble code from scratch, without relying on any ensemble-related libraries.

Before evaluating the performance, we will first conduct a code plagiarism check. The performance of each implementation will be assessed using our private test dataset. For evaluation, we will use accuracy and F1 score metrics. Additionally, we will measure the running time of each implementation.

For this assignment, each team must submit a single Python file (.py) along with a Colab notebook link. The notebook should include clearly labelled headings for each cell for the execution. Afterwards, create a plot to visualize the performance comparison of all your implementations, including the ensemble model. Add this plot and associated discussion in the report prepared for part A.

We'll use MS Teams to accept the submissions. Only one member from each team should submit the assignment deliverables.